Streamlining the “Producer/Archive” Interface: Mechanisms to Reduce Delays in Ingest and Release of Social Science Data

Jinfang Niu
Margaret Hedstrom
University of Michigan

Outline

- Background
- Research questions
- Methodology
- Findings & discussions

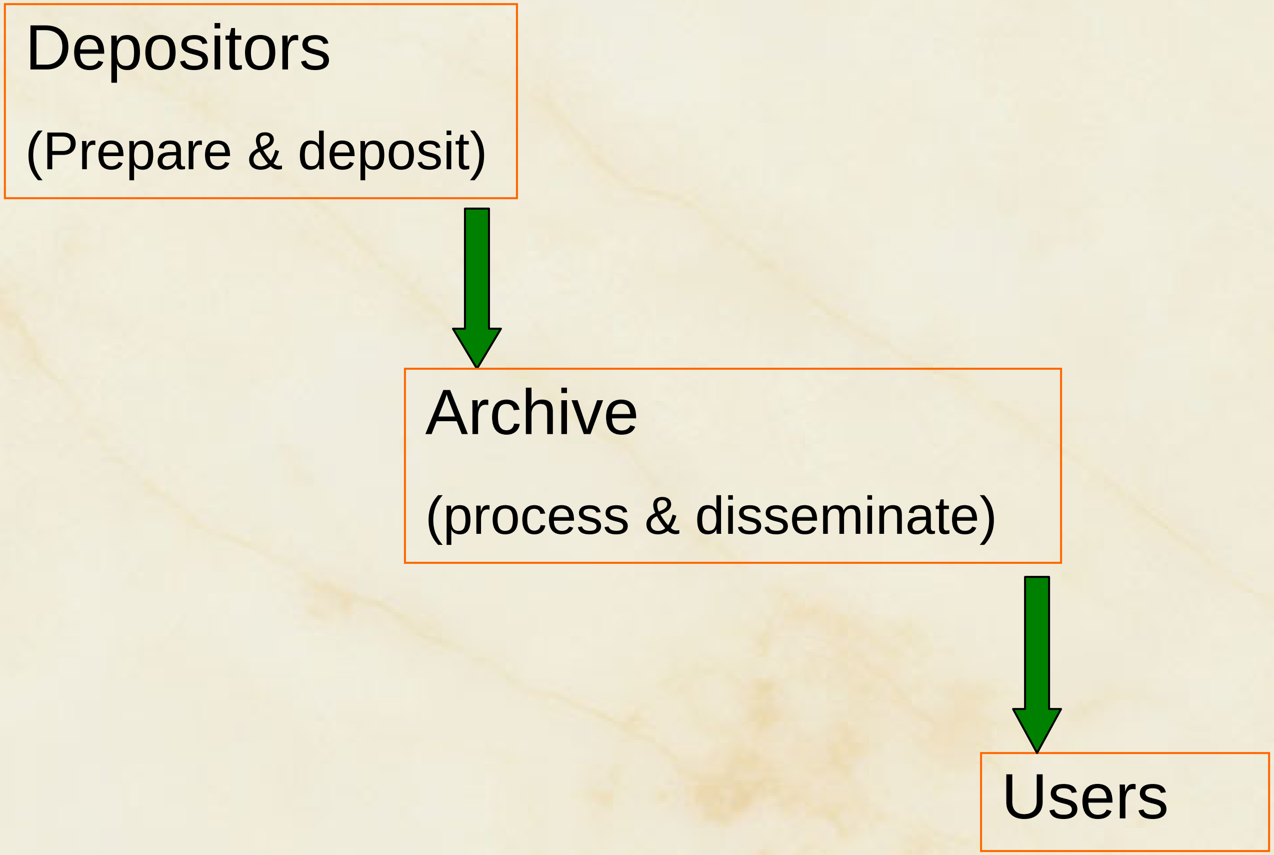
Data sharing is a growing concern

- Government policies
  - OECD
  - FOIA
- Funding agencies
  - NIH, NSF, ESRC
- Journals

Sharing data through data archives

- Long-term preservation
- Data archivists help both depositors and users
- Make it possible for meta-analysis
- Improve the visibility and possibly the citation rate of data.

Sharing model through data archives



Depositors

(Prepare & deposit)

Archive

(process & disseminate)

Users

Good data archiving practice

- Data producers
  - Deposit in the appropriate data archive
  - Prepare data well
  - **Deposit in a timely manner**

- Data archive
  - Processes and releases data for public use as soon as possible

- Users
  - Gain access to deposited data as soon as possible
  - Use data without too many difficulties

Research questions

- Do producers deposit data in a timely manner?
- How quickly does the archive release data to the public?
- What causes the delays?
- How to improve the situation?

Methodology

- Analysis of delays (n = 184 data sets)
  - Deposit Delays
  - Processing Delays
- Causes
  - Submission and processing procedures
  - Incentive issues for depositors
- Proposed Solutions

Delays

- Deposit delay

  the number of days between the date a grant was closed and the date that the data archive received the data.

- Processing delay

  the number of days between the date when the data arrive at the archive and the date when the data is released to the public.

Delays (in days)

	Mean	Median	Min	Max
Deposit delay	767	664	-27	2630
Processing delay	355	276	20	1187
Total	1160	1122	263	2846

Causes of deposit delay

- Two-step submission procedure

  Data depositor → funding agency → archive

- No clear timeline for deposit
- No effective incentive mechanisms

Processing Delay vs. Actual Processing Time (in days)

	Mean	Median	min	max
Processing time	10	7.5	1	45
Processing delay	355	276	20	1187

Causes of processing delays

- Depositors submit incomplete data and documentation
- Depositors review the processed data.
- Funding agency delays transmission of final reports to the data archive
- Extremely large data sets require more time to process and delay processing of other data sets in the queue

Proposed solutions - 1: Streamline submission process

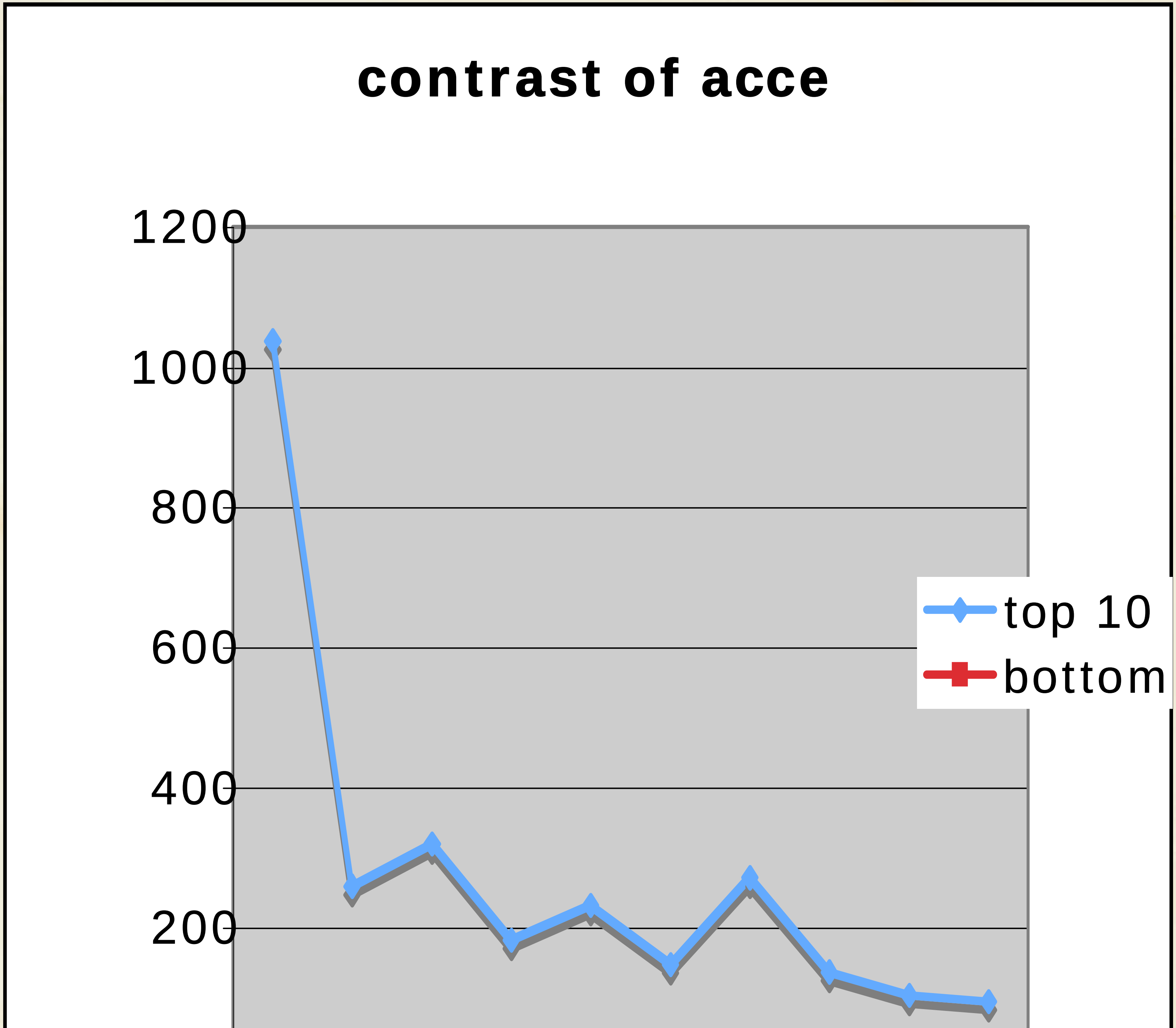
- Change the data submission procedure
- Stipulate a clear timeline for deposit
- Improve the awareness and availability of documentation guidelines

Proposed solutions - 2: Incentives

- Punishment
  - Coercive and uniform
  - Pros and cons:
    - Makes all data accessible to the public.
    - All data producers have to prepare and deposit data to avoid punishment even if their data sets are not likely to be used.

Cumulative and Average Monthly Access Rates for the 10 Most Frequently (Top 10) and 10 Least Frequently (Bottom 10) Accessed Data Sets.

Cumulative		Average (per month)	
Top 10	Bottom 10	Top 10	Bottom 10
5185	27	1037	3
2345	25	260	2.8
2234	24	319	2.7
1651	24	183	2.7
1623	23	232	2.6
1328	21	148	2.3
1267	20	271	2.2
1229	16	137	1.8
932	14	104	1.6
851	11	95	1.2



contrast of acce

Proposed solutions - 2: Incentives

- Rewards
  - Inducive and selective
  - Pros and cons:
    - Related to the actual use of data
    - Difficult to anticipate actual use.
    - Not all data are accessible to the public

Future research

- Exploration of appropriate punishment & reward mechanisms
  - Proposed mechanisms
    - Hold back a portion of grant funding
    - Make future funding contingent on data deposit
    - Citation of data
    - Include data deposit in performance evaluation

Thanks!

This research was supported by the National Science Foundation (NSF Award Number IIS-0456022) as part of a larger project entitled “Incentives for data producers to created archive-ready data sets.”